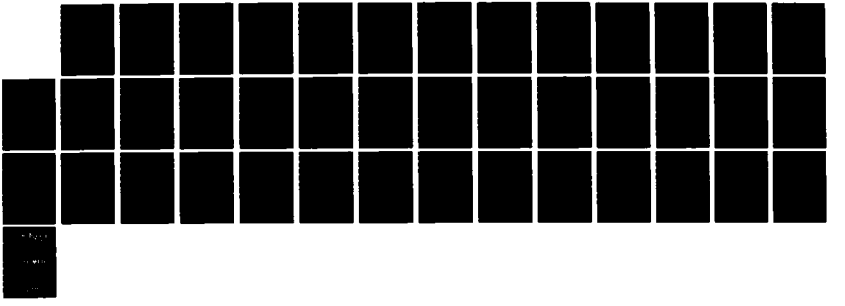
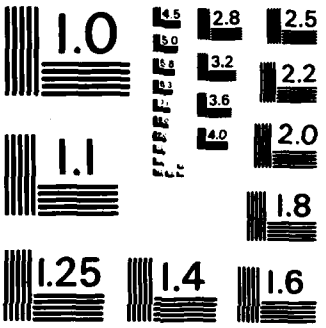
AD-A162 504 GEONAMES PROCESSING SYSTEM FUNCTIONAL DESIGN SPECIFICATION VOLUME 2 GEOGR..(U) NAVAL OCEAN RESEARCH AND DEVELOPMENT ACTIVITY NSTL STATION MS.. 1/1

UNCLASSIFIED G LANGRAN ET AL. MAR 85 NORDA-99-VOL-2 F/G 9/2 NL





Naval Ocean Research and Development Activity
March 1985 **Report 99**

Geonames Processing System Functional Design Specification

Volume 2: Geographic Names Data Base

Gail Langran

Mapping, Charting, and Geodesy Division
Ocean Science Directorate

Allen Barnes
Steven Miller

Planning Systems, Inc.
McLean, Virginia

FOREWORD

DMA has recognized a need for digital procedures to store, retrieve, and edit geographic names and to prepare names data for product generation. DMA's stated goal is a 50–100 million name digital data base with subsystems to capture names data, edit and format data, and prepare names overlays for maps. NORDA began a geonames processing system design study late in FY82. This report is one of a five-volume series of reports that describe the functional design of the digital geographic names processing system.

R. P. Onorati, Captain, USN
Commanding Officer, NORDA

This NORDA Report was prepared to meet style requirements of the sponsor.

EXECUTIVE SUMMARY

In FY82, the Pattern Analysis Branch, Mapping, Charting and Geodesy Division of the Naval Ocean Research and Development Activity (NORDA) began a subtask for the Defense Mapping Agency (DMA) entitled, "Advanced Type Placement and Geonames Database System Development." This effort will develop systems to address four interrelated aspects of computer-assisted geographic names processing as follows.

- Data Capture: digital capture of names and named feature information from analog sources such as maps, gazetteers, and other data sources.

- Data Management: development or adaptation of a Data Base Management System for a very large product-independent set of world geographic names and their descriptors. This data base will support a variety of DMA products including maps, charts, and gazetteers.

- Data Manipulation and Editing: in support of toponymic research, advanced word processing for text containing diacritics and special symbols, document formatting, data file searching, and statistics generation.

- Product Generation: digital text placement on maps, gazetteers, and other DMA products with the associated data selection, formatting, scaling, and type assignment.

This Geonames Processing System subtask is scheduled for performance during FY82–FY89. During the first year (FY82) an initial Comprehensive Coordination Plan (CCP) was generated for the technical description of the above automated names capability (NORDA Technical Note 189). The second stage of planning built on the CCP and DMA responses to the CCP to generate the present five volume set of Functional Design Specifications, one for each subsystem and one to describe requirements that are mutual to all four subsystems.

This report states the functional requirements and specifications of a geographic names data base. Data administration, the user interface, data entities, external interfaces, and hardware and performance requirements are discussed.

ACKNOWLEDGMENTS

This work was sponsored by DMA under Program Element 64701B, with subtask title, "Geonames Processing System." Mr. Dennis Franklin and Lt. Col. Tom Baybrook, both of DMAHQ/STT, shared project management duties during the writing of this report. Their help in communicating with DMA's production centers and providing information on DMA production methods was instrumental to this functional design. Dr. Don Durham, head of NORDA's Mapping, Charting, and Geodesy (MC&G) Division, and Dr. Charles Walker, head of the MC&G Division's Pattern Analysis Branch, contributed valuable advice and assistance.

TABLE OF CONTENTS

FIGURES

TABLES

INTRODUCTION

a. Organizations

Defense Mapping Agency Headquarters (DMAHQ)
U.S. Naval Observatory
Washington, D.C.

Defense Mapping Agency Hydrographic/Topographic Center (DMAHTC)
6500 Brookes Lane
Washington, D.C.

Defense Mapping Agency Aerospace Center (DMAAC)
3200 South Second St.
St. Louis, Missouri

b. Scope

The purpose of this report is to describe system attributes, serving as a basis for mutual understanding between the user and the developer.

The Geonames Processing Subsystems are often referred to in this report by their acronyms: ASP (Advanced Symbol Processing); ATP (Advanced Type Placement); GNDB (Geographic Names Data Base); and AADES (Automated Alphanumeric Data Entry System).

c. Background

In FY82 the Pattern Analysis Branch, Mapping, Charting, and Geodesy Division of the Naval Ocean Research and Development Activity (NORDA) began a subtask for the Defense Mapping Agency (DMA) entitled "Advanced Type Placement and Geonames Data Base System Development," a project encompassing the digital capture, storage, edit, and display of geographic names. The subtask in its current form is an amalgamation of four previous DMA requirements for independent development of a geographic names data base, a system for high-volume geographic names data capture, advanced word and symbol processing, and automated type placement for maps (see Appendix B for DMA's original requirement statements). A Comprehensive Coordination Plan was submitted by NORDA as a preliminary definition of the overall Geonames Processing System subtasks and their interfaces.

d. Description

The complete Geonames Processing System is comprised of four components (Fig. i-1).

- The Automated Alphanumeric Data Entry System provides a means of high-volume geographic names data capture. World geonames with their corresponding locations and attributes will be captured from both tabular and map/chart sources using raster scan and optical character reading technologies. AADES converts alphanumeric data into computer-readable form with a 99% accuracy rate. It requires minimum operator intervention, pro-

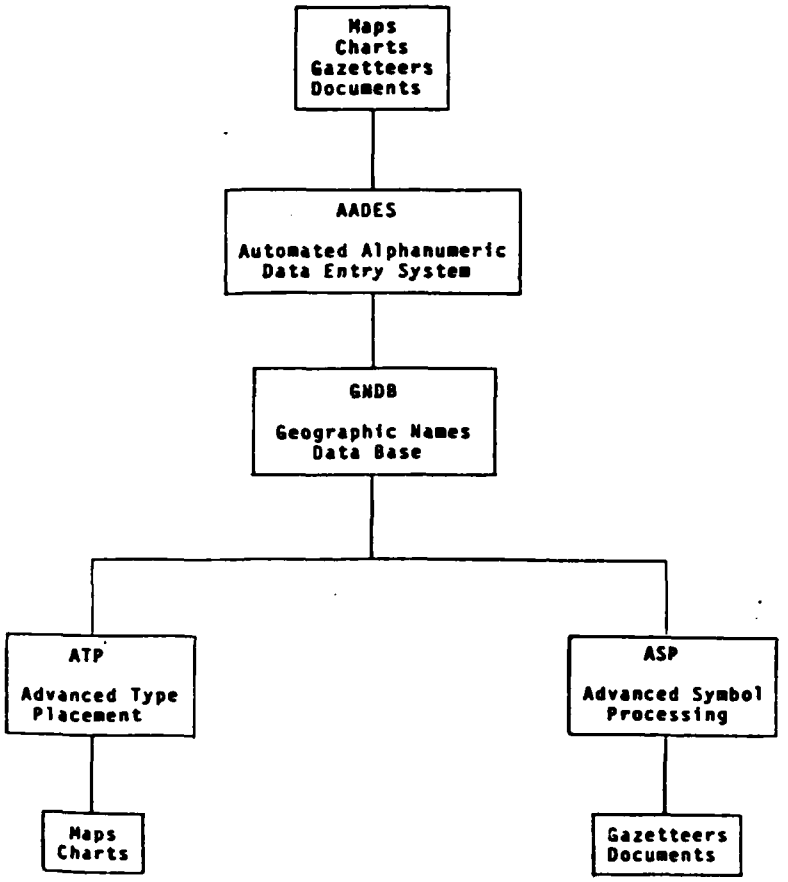


Figure i-1. Geonames Processing System overview

vides automated error checking, and results in clean data files for supervised merging with the GNDB.

- Geographic Names Data Base stores world geonames and their descriptors in non-product-oriented files. It provides extensive query capabilities to support data base updates, chart and gazetteer compilation, and toponymic research. The GNDB's ultimate size will be 50–100 million geonames.

- Advanced Symbol Processing. International geonames comprised of diacritics and special symbols require specialized hardware and software for access, manipulation, and editing. ASP provides alphanumeric edit and display of world geonames and advanced word processing capabilities such as sorting, searching, and formatting.

- Advanced Type Placement automates the production of map names overlays by exploiting electronic display technology and the rule-based nature of cartographic names placement. ATP includes automated utilities for names selection, type composition, type placement, virtual map display, and interactive graphic edit.

The Geonames Processing System responds to a major need: it will integrate DMA's names processing tasks into the digital map production pipeline and coordinate all geonames processing ac-

tivities. Obvious benefits are increased production rates and lowered costs. Overall accuracy and coverage should also improve with the increased efficiency of such a system. Helpful utilities will raise toponymic researchers' productivity levels. Further automation will be easier to implement once the process is converted from manual to digital.

This Functional Design Specification addresses the development of Advanced Symbol Processing for international geonames (ASP). Overall processing flow is outlined in Section 8. It is suggested that the reader turn to this section before continuing.

e. Applicable Documents

The following references provide a summary of the basis for the Geonames Processing Subtask development.

- "Advanced Type Placement and Geonames Database: Comprehensive Coordination Plan." NORDA Technical Note 189, January 1983.
- "A Prototype Geographic Names Input Station for the Defense Mapping Agency." Paper presented at Auto Carto IV by D.R. Caldwell and D.E. Strife, September 1982.
- "Names Type File System." Consulting Report for the U.S.A.E.T.L Project #POO13, April 1983.
- "Development of an Automated Cartographic Capability." The Final Report of the Automated Cartography Task Force, Defense Mapping Agency Hydrographic/Topographic Center, April 1982.
- "The Feasibility of Establishing an Automated Chart Production Process." The Defense Mapping Agency Hydrographic/Topographic Center, October 1982.

f. Limitations

The individual subsystem descriptions are functional and not physical definitions, i.e.:

- a given function required by a given subsystem may not be performed upon the hardware logically associated with the subsystem,
- one software module may serve several of the subsystem functional requirements.

Thus, the functions and data sets specified in this five-volume set of design specifications are described somewhat redundantly to fully define each subsystem regardless of its interplay with other subsystems. Physical (hardware and software) synthesis will be accomplished and described by the Implementation Plan at a later date.

1.0 GEOGRAPHIC NAMES DATA BASE OVERVIEW

1.1 Existing Methods and Procedures

Current DMA operations for producing gazetteers and maps are labor intensive in regard to the handling of geographic names. For gazetteer production, the primary reference file is the Foreign Place Names File (FPNF), which consists of about 4.5 million names and associated data on index cards. This file conforms with the Board on Geographic Names approved spellings. For map production, some geonames may come from the FPNF, but many are taken from existing maps. Hence, there is an analog map data base for those geonames not among the 4.5 million in the FPNF. Forms called Names Data Records are used to transfer names from the analog maps to the digital Multiset III system. The Multiset III, an advanced typesetting system, is used to prepare the names for products. As part of this process, the names are captured in digital form on diskettes.

1.2 Deficiencies and Drawbacks

DMA updates gazetteers at the rate of about 10 per year. Because of the large number of gazetteers maintained, some are very out of date (10–25 years old). Since gazetteers are produced from the FPNF while geonames for map products are often taken from other maps, DMA gazetteers and maps do not always agree. A common geonames file for both toponymic and cartographic applications should increase standardization and reduce processing requirements, resulting in a savings of both time and money.

Since the FPNF is on index cards, queries are handled manually, which is very time-consuming for some types of questions. The 4.5 million names in the FPNF roughly correspond to the geonames on 1:250,000 scale maps. Names that appear only on maps of large scale (1:100,000 or 1:50,000) are generally not in the FPNF.

1.3 Proposed Methods and Procedures

A subsystem of the Geonames Processing System, the GNDB will handle the data banking, storage, and retrieval of geonames and their associated attributes in an all-digital environment. This will provide a single, controlled repository for geonames data within DMA and, as part of a computer-assisted cartographic system, will increase production throughput. The file will consist of about 60 million geographic names and their attributes, thus serving both toponymic and cartographic products. The GNDB will receive batch inputs from AADES, and will allow authorized operators to modify the data base from interactive ASP workstations. Data sets will be compiled from the GNDB in support of gazetteer and map production. The GNDB will reside at DMAHTC and all updates will be made by DMAHTC personnel. The GNDB will output Names Data Files to DMAHTC users via terminals and to DMAAC users via batch transmission (e.g., magnetic tape or data line).

1.4 Ordering of this Functional Design Specification

Section 2 discusses GNDB data administration. Requirements to support system operators (in both interactive and batch modes) are given in Section 3. GNDB functions are discussed in Section 4 and Section 5 describes data entities and data sets. Section 6 deals with GNDB interfaces to other subsystems and external systems. Architectures and data structures are examined in Section 7. Section 8 lists the basic assumptions of this design. Requirements for performance and hardware are discussed in Sections 9 and 10, respectively.

2.0 DATA ADMINISTRATION

The Geographic Names Data Base is a complex relationship of people, software, and computer equipment. Critical success factors for meeting GNDB goals include the capabilities of the data base management software (DBMS), the operating system, and the data base administrator. Performance requirements of the DBMS are addressed in Section 9.

2.1 Operating System (OS) Capabilities

Minimal OS requirements include supporting multiple users in both batch and interactive modes, multi-tasking facilities for up to 20 simultaneous processing requests (i.e., 20 terminals), common memory workspace areas for users to share, and flexible and efficient interfaces to the data base management software. Several GNDB functions could be supported by either the DBMS or the OS (e.g., data security, audit trail activities). To optimize the potential synergism of the OS/DBMS relationship, OS system alternatives should be considered in conjunction with data base management software evaluations.

2.2 Security Considerations

2.2.1 Security Classification

All data entities planned for inclusion in the GNDB are unclassified. The GNDB software is also unclassified. The GNDB interfaces with ASP, and through it with AADES and ATP. These systems are all unclassified. If data from the GNDB is to be transmitted to a classified environment, media such as magnetic tape or floppy disk may be used.

DMA has indicated that the aggregate of geographic names and their attributes forms an unclassified unit, and thus the GNDB should be unclassified. However, DMA has expressed concern that this system may at some time be deemed classified. There are three potential areas of concern.

- The aggregate of names may indicate (by density or completeness of coverage) regions of high interest to U.S. targeting operations.
- Certain data base queries may indicate (by the area or the strip selected) planned approaches to a target by bombers, cruise missiles, etc.
- One may wish to merge unclassified geoname data with classified data in support of a classified product. Such a merger could be supported by ASP. However, DMA may wish to place classified data into a separate data base similar to the GNDB for ease of use. Such classified data may consist of classified attributes associated with otherwise unclassified geonames records, or it may consist of entire classified geonames records.

This functional design assumes that the GNDB is unclassified, but allows the option of building a parallel classified data base (much smaller than the GNDB) should such action be warranted in the future.

2.2.2 Access

The unclassified GNDB will require access controls. These controls will ensure that

- only authorized users may gain access to the system;
- only an authorized few may change the data base;

- private data files are modified or purged only by their creator (or by the system administrator, in the event of problems);

- access to system files and management information statistics is restricted to system personnel, and changes to auxiliary files (such as the Coordinate Limits File) are made only with the approval of system personnel.

If a classified parallel base is built, it will be subject to the same access controls, as well as those imposed by security regulations.

2.3 Data Base Preservation

The design of the Geonames Data Base must incorporate fault-tolerant principles to protect the integrity of the data. DBMS data preservation software is addressed in Sections 9.1-9.3; the role of data administration in data preservation is discussed here.

The system must have backup facilities that copy data base contents to removable media (e.g., tape) for secure storage in a location removed from the GNDB to minimize the consequences of catastrophic events (fire, flood, etc.). Once the system becomes operational, the cost of reconstructing the data base will outweigh hardware replacement costs. Backup procedures should require minimal operator involvement; provisions for generating multiple copies simultaneously (to allow operators to keep current backup files on hand for noncatastrophic recoveries) is desirable.

The frequency of generating backups will be dependent on data base volatility. During initial loading frequent backups (several times a week) will be required. The mature data base should require only monthly backups. Depending on the system architecture (see Section 7.1), backups of data base segments or area-specific data bases may be required rather than full-system backups. Both the latest backup copy (the "father") and the previous copy ("grandfather") should be maintained.

With most data base updates resulting from batch processing, system failures should trigger roll-backs of data base contents to the state prior to the last batch run and generate appropriate error diagnostic messages. Automatic restart of the run in which the error occurred (for transmission and software errors) would minimize operator intervention requirements. Hardware failures (e.g., disk crashes or nonrecoverable memory errors) will require operator intervention to recreate the data base from backup copies, including all update jobs run since the time of the backups. Small data sets that were manually entered into the quality control process (see Section 4.2) should be recorded as part of the current batch update to minimize manual data re-entry following failures, and to facilitate the data redundancy checks that are part of input processing. Isolating input files from data base update routines (which actually modify the sorted data) will help preserve the consistency of data outputs and reduce average restart times.

When choosing a hardware configuration, fault-tolerant principles to prevent total system failure resulting from malfunctioning components should be considered. Redundant hardware and automatic switching circuits are desirable for critical system elements. An operating system that detects hardware failures is desirable. System power supply protection must also be addressed in evaluating hardware alternatives.

2.4 Data Base Administration (DBA)

An individual or an executive committee can administer the GNDB. The DBA controls data base information content and preserves system usefulness and longevity by establishing policies and procedures. DBA is distinct from production management in its long-range, nonterritorial perspective on managing information resources. The ultimate value derived from the data base investment is largely dependent on successful implementation of DBA functions.

Primary DBA responsibilities include establishing and policing standards for data size, format, and usage; administering detailed system documentation; and coordinating user needs in light of current system capabilities and resource development objectives. The DBA arbitrates disputes over information ownership, access priorities, and design issues. Arbitration requires clear delineation of DBA authority and responsibility. Other DBA functions include software acquisition, performance management of both the DBMS and system procedures, and data security administration.

The scope of the Geonames Processing System and the anticipated volume of GNDB data suggest that the DBA role be established early in the design phase (Task 4 in Reference A). Top-level support of this function will facilitate system implementation and integration with other Geonames Processing subsystems and enhance both system performance and cost/efficiency.

2.5 Impact of GNDB Requirements on DMA

The GNDB initially will need at least two people to fulfill DBA functions; two system administrators to establish input quality control, output production standards, and maintenance procedures; and two applications programmers to develop and enhance DBMS software interfaces. As the system matures, the size of this support staff group can be reduced.

Physical space requirements for the GNDB are estimated as

- 6 administrative/support @ 120 sq ft = 720
- 20 operator workstations @ 30 sq ft = 600
- hardware and production equipment (4 areas @ 16 ft x 20 ft) = 1280

2600 sq ft

While workstations could be distributed within a building through local-area networking, administrative offices should be located close to system operations to facilitate administrative tasks. The impact of personnel and space requirements on integrating the GNDB with current DMA facilities must be addressed when considering alternate system designs.

3.0 USER INTERFACE

The Geonames Data Base management system must provide users with two classes of data access: read/write and read-only. Centralized control of read/write operations is required to preserve data consistency and integrity. Modifications other than routine loading are reviewed by the appropriate authority before execution. The majority of system usage will be read-only, i.e., queries or file compilation.

3.1 User-Invoked Functions

The DBMS operating system should maintain clocked logs on system usage. Coded access (a password system) is desirable to facilitate system administration and control, with priorities (if required) assigned by the DBA. Coordination between user groups and communications among users would benefit from a message (''mail'') facility.

Users responsible for data base updates invoke application software to filter input files through interactive quality control checks prior to being loaded into the data base. Actual data base loading, however, is in batch during slack periods to avoid degrading interactive performance. During quality control procedures the analyst generates a file of questionable data records for subsequent investigation and correction. Accurate records of this process should be maintained (either manually or by the system software).

A query language that supports multiple selection criteria (both by inclusion and exclusion) is required to support interactive data base queries. Facilities to compile and store query sequences for re-use on standardized products are desirable. If feasible, a system for returning a count of data base entities that meet the query selection criteria without having to search through the data should be developed to avoid repeated trial-and-error queries. Data manipulation requirements are addressed in Section 4.4.

All queries will generate or add to user working files by searching the data base (read-only), sorting by user-specified criteria, and copying the results to a storage area (which may be a file or terminal memory). Facilities for editing (inserting, deleting, and modifying) and resorting the resulting user data are required, with provisions for subsequent file storage and hardcopy output of the results. A library of user files is expected to evolve, requiring file management software to control cataloging and access procedures.

3.2 Terminal Requirements

The GNDB will be accessed by interactive users through ASP terminals. The terminals must be capable of handling Latin letters (in upper and lower case), diacritics on the Latin letters (there may be more than one diacritic associated with a letter), numerals, and the usual characters (e.g., dash, space, period, comma).

The ability to display non-Latin alphabets such as Greek, Cyrillic, or Korean is not required. Such capabilities, however, may be desirable, and considerations concerning their support are discussed in Volume 5, Appendix C of this report series.

Bit-mapped graphics are not required for GNDB displays, unless non-Roman script is to be supported. However, pixel-addressable CRTs are desirable to provide rudimentary geographic feedback to quality-control monitors on input and to improve format and display capabilities on output. CRT screen dumps to a non-impact graphics printer provide a rapid and efficient means of generating working hardcopies.

The system architecture selected (Section 7) will have a major impact on the data processing capabilities and memory storage requirements of the CRTs. With a single mainframe architecture, computer memory files could be manipulated by "dumb" terminals; a distributed architecture requires more intelligent terminals. Minimum memory requirements to support downloading of data base subsets for subsequent processing is also a function of the system architecture selected.

3.3 Hardcopy Requirements for Users

Interactive users require hardcopy support. Printers must print the standard 132 characters, and diacritics as well. The printer system currently used with the NIS has diacritics, but is too slow for a production environment.

A plotter is useful for quality-control reviews. Because the plotter is used as an analyst aid, not for product generation, high resolution is not required.

4.0 GNDB FUNCTIONS

4.1 Processing Flow

The GNDB will receive data inputs from a variety of sources. Its output supports queries and product generation. A description of GNDB processes follows.

- User takes an input file in interface format and executes the following quality control and update modules.

  - Check new data against the data base for consistency.
  - If the new data is accepted, put it into an update file for addition to the data base.
  - If the new data improves upon existing data, put it into the update file to change the data base.
  - If the new data conflicts with existing data, put it in an exceptions file for manual attention.

  The update file mentioned here may either be a physical file that will be processed in batch (e.g., overnight), or a DBMS update buffer. (The latter approach requires DBMS concurrency control.)

- Toponymist edits the exceptions file, correcting or deleting records. When the toponymist believes the problems are resolved, the file (in interface format) goes through quality control procedures.

- Toponymists working on interactive ASP workstations make changes that are batched together for data base update that night. Toponymists, like any users, may modify their working files immediately; but for data base control, all updates are performed in batch either at night or by authority of the DBA.

- Users compile Names Data Files for product generation. Files are reviewed at ASP workstations or in hardcopy form.

- Names data may be modified during product generation if errors are discovered. Such modifications are written to a Data Base Update File for entry to the GNDB.

4.2 Input Processing

There are four major types of GNDB inputs. These are listed below, along with their processing requirements.

- Batch updates from AADES. These are entered via ASP, in interface format, and passed directly to the quality control modules (see 4.3).

- Multiset III tapes, foreign tapes, other agency tapes. Geographic names with varying attributes are available in digital form from varied sources. These will be converted to 9-track tape, then input to the GNDB via ASP. A software module will be needed to convert data from their existing format to the interface format and to fill any required fields (e.g., feature designator) that may be missing. Once in interface format with all required fields filled, files are treated as input files from other Geonames Processing Subsystems.

  Geonames data processed on DMAHTC's current Names Input Station (NIS) will be treated the same way. It is anticipated that by the time the GNDB is operationally loaded, the

geonames data available from the Multiset III and the NIS will make it unnecessary to load the old 7-track gazetteer tapes (see Ref. a).

- User changes and additions. Authorized users at interactive terminals may modify the data base. These modifications are not done immediately but are batched. The updates are received via the ASP and passed to the quality control modules.

- Loading and changing auxiliary files. A number of auxiliary files are used by the GNDB, such as the Coordinate Limits, Area Code Boundary, and Map Sheet Boundaries Files. The interface format is neither appropriate for loading data into these files nor for updating them. Software for loading an entire static file, or modifying a static file, will be required. Input processing consists of recognizing the input, reformatting if necessary, and forwarding it to the DBMS update modules.

4.3 Quality Control of Input

Quality control may be divided into three types of tests (described further in Volume 4 of this series).

- Range tests. All input fields are subjected to range tests to identify absurd values (e.g., latitude greater than 90°, negative population, unrecognizable feature designator). Positions will be checked against the coordinate limits file to ensure that they are within the identified country.

- Duplication tests. Names will be checked against the data base to see if the feature is already in the base. Checks will be made interactively for nearby features with similar spellings. If a duplication exists and the data is inconsistent (e.g., positional difference is not within the positional accuracies) then the conflict must be resolved. The question of whether two features with nearby (but not identical) positions and similar (but not identical) names are actually distinct should be resolved by a human. A later implementation may have the initial checking performed by the computer.

- Resolution tests. AADES produces digital data from analog media. When overlapping map sheets are digitized, one may expect differences to occur between two sets of digital data referring to the same feature. The digitized positions should be close, but will not be identical. The feature attribute, such as population (determined by the size of symbol or size of letters), may differ due to age difference of the maps. When two sets of data are available on a single feature, and they are inconsistent, then the analyst must decide which set to keep in the data base. This decision is made based on geodetic control of the input maps, map scales, positional accuracy, and data source and date.

4.4 Data Base Manipulation

Apart from input procedures and system statistics, users will be concerned with extracting information from the data base. Two major types of user interfaces are expected—simple query extraction, and compilation and execution of applications involving query sequences, edit facilities, and tools to assist in formatting outputs.

Common access parameters will include geonames, geopolitical entity types, geographic area coordinates, and feature designator types. Facilities to select subsets of data sets accessed on primary parameters are required, including logical associations among attributes and toponymic elements, geographic proximity to other entities, and data sources.

Query facilities should provide programmable, transparent interfaces to applications software for requests not supported by the query syntax. For example, a query to locate entity names by lexical similarity should automatically invoke the appropriate processing routines (such as the name variant function described in Section 4.5.4). The query language syntax should provide for two levels of user sophistication, with menus or help sequences for novices, and compact notation for experienced operators.

The data dictionary (a feature common to most commercial DBMS software) provides a catalog of data specifications and relations present in the data base. This should be made available online to users for reference and assistance in querying the data base. Meaningful error messages explaining inappropriate requests are required rather than cryptic responses ("no records satisfy your request") or no response at all. The DBMS must support all basic data manipulation operations, including insert, edit, delete, access by query, retrieval, and sorting and formatting by user-specified parameters. File management facilities that provide automatic and optimized data base navigation are required. Pagination, menu selection and design, and graphics primitives are desirable management reporting facilities to support operator efficiency and ease of use.

4.5 Applications Software

GNDB operators will interface with the data base management software through DBMS-supported query language expressions, or through applications programs written in procedural languages (e.g., Fortran, Pascal) designed to support specific system functions. Authorized users and programmers must have provisions for concatenating queries and storing them (preferably online) for reuse. Users should be able to selectively direct query results to file storage, and buffer working area memory or intelligent terminal memory. Provisions for invoking query sequences from inside edit and format routines written in procedural languages are desirable.

The number of modules in the category of "applications software" will grow with time. As a minimum, the following functions will be addressed by applications software.

4.5.1 Conversion Functions

Three functions fall under this heading. Each takes a position (latitude, longitude) and converts it to a UTM grid, a Map Sheet number, or an Area code for gazetteers. UTM grid conversion should be performed using a mathematical algorithm. The conversion to map sheet number is required for scales of 1:50,000, 1:100,000, 1:200,000, 1:250,000, 1:500,000, 1:1,000,000, 1:2,000,000 and 1:5,000,000. This function will use the map sheet files described in Section 5. The conversion to area code cannot be performed by a simple mathematical algorithm. An auxiliary table indicating area code boundaries should be used to implement this function.

4.5.2 Product Inclusion Functions

Product inclusion functions take a position and a set of product limits and determine whether or not the position falls on that product. For maps, projection is computed first.

4.5.3 Name Reduction Function

This function will not be invoked directly by the user, but will be used by other procedures such as quality control, data base update, query, and sort. The name reduction function takes a geographic name and constructs a reduced text string. To do this, it converts all letters to upper case, removes all diacritics, removes recognized prefixes and suffixes, removes spaces and dashes, and eliminates abbreviations. There are two modes of operation: the difference is in the removal of diacritics. In one

mode (used by query and sorting), the diacritics are simply dropped. In the other mode (used by the names variant function), all diacritics are dropped. For specific diacritics in specific languages, a diacritic and its vowel may be replaced by a dipthong. The reduced text string is intended strictly for internal use; it should never be accessible to the user since it is not an actual name.

To illustrate the function of this module, the following five examples give pairs of names that have identical reduced text strings. Thus, the software can recognize the basic similarity of each pair. These include

- Los Banos and Los Baños;
- Bay Port and Bayport:
- Weston-super-mare and Weston Super Mare;
- Al Basrah and Basrah;
- Mt. Zion and Mount Zion.

4.5.4 Output Formatting Functions

Compilations from the data base for gazetteer or map products must be placed in Names Data File format (Section 5). Other retrievals must be placed in appropriate format for display, printer, or plotter output. Files for systems external to the Geonames Processing System must have DMAFF headers.

4.5.5 Encoding Functions

These functions are invoked automatically by the DBMS; they cannot be called by the users. The data base contains a number of data entities that, as far as the user is concerned, have alphanumeric values taken from a fixed set. An example of such a data entity is the feature designator; it must conform to the set of legal values given in the data dictionary to ensure that AADES, GNDB, ASP, ATP and the users recognize the designator used and that all agree on a common definition. Within the GNDB itself, however, these alphanumeric fields may be translated into numeric codes to decrease the amount of storage needed. This encoding process is part of many DBMS packages; but since it is not universal, it is considered here as a specific applications process.

The encoding process is actually an aspect of implementation, not of functional design. However, due to the large number of geoname entries, failure to use encoding (or similar storage reduction technique) will require an additional 1.5 gigabytes in the Audit Trail Data File alone (see Section 6.2.5). The large difference in storage requirements attributable to encoding is of sufficient importance that this function is considered a design requirement.

4.5.6 Date Function

An assumption is made that the current date is kept somewhere in the machine (usually by the operating system) and that other software has read-only access to this date. This function allows the GNDB to access the date to fill date entries in the Audit Trail Data File.

4.6 Output Generation

GNDB outputs are either quick queries to explore data base contents or more complex query sequences (or access applications software) invoked to support file compilations. Data retrieved from quick queries will require terminal display and dumps to a printer, while outputs from more elaborate requests will generally require interactive editing and formatting before subsequent file storage or hardcopy dumps. Graphic displays on terminals or digital plotters should be supported.

Facilities to compare, contrast, combine, sort, and separate outputs in conjunction with stored files (and files with other files) are required, along with provisions for generating tape copies and transform-

ing files electronically. These functions may be supported by the operating system, data base management software, or applications software. Stored data sets will require restructuring to Names Data File format for efficient communication with ASP and ATP.

4.7 Management Information Statistics

Information on current system functions, changes in data volume, and measures of DBMS efficiency over time are required by the DBA to monitor system performance and to aid in planning system resource allocations. Identification of on-line users, current processing functions, and system hardware allocations should be accessible interactively. Information on volumes (number of elements), access frequencies, and measures of update volatility of various data base components should be made available to system operators.

Tracking trends in data base growth and retrieval efficiency will allow the DBA to trade off storage allocations, access algorithms, and user demands in tuning the DBMS for optimal performance. Advance warning of processing bottlenecks will facilitate system adaptability to change. Suggested measures of retrieval efficiency are indicated in Section 9.6.

Audit trail functions, including records to identify the user who inserted or modified a data element, when this occurred, and references to special circumstances of the event are required to support some types of toponymic inquiries. This information is not required to be stored online or available for general access, and occasional requests of this type are expected to be processed in batch. System software to automatically capture audit information during data base loading is desirable.

5.0 DATA ENTITIES AND DATA SETS

5.1 Data Entities

- Country: alphanumeric, 40 characters maximum. Can be alias name (e.g., USSR or CCCP for the full name), but the alias must be in the dictionary and identifiable by the GNDB or the record will be rejected. Each country must have a unique official country name.

- Data Source Name: alphanumeric, 10 characters maximum. Up to six data sources per name may be used. If a data source does not match any known to the GNDB, the GNDB will add it to the list of legal data source names. Thus, aliases should never be used here for this would adversely affect retrievals based on data source names. DMA should decide on an appropriate set of data source names (to avoid alias problems) and place these in the dictionary.

- Date of Data Source: numeric string, 7 digits. The date will be stored as "ddmmyyy" where dd represents the day (0 through 31, 0 representing an unknown day), mm represents the month (0 through 12, 0 representing an unknown month) and yyy representing the last 3 digits of the year. Three digits are not necessary for the year, but they make it easier to handle any nineteenth century data sources, and will allow the GNDB to handle the transition to the twenty-first century without modifications.

- Date of Data Capture: numeric string, 7 digits. Same format as date of data source, except that the day and month elements are mandatory—the system should prevent "unknown" entries. This date should be machine generated by the AADES.

- Date of Last Update to Record: numeric string, 7 digits.

- Comment field: alphanumeric, 256 bytes. For research notes.

- Geographic Name: alphanumeric, 40 characters with provision for overflow. Name is in Latin alphabet, upper and lower case, with diacritics. The name may contain hyphens, spaces, or periods. The name should appear as on the source material. The GNDB will convert to upper and lower case, expand abbreviations, etc., if needed. Diacritics will be represented as 1-byte codes, with an indication of which characters they append.

- Type of Romanization: alphanumeric, 6 characters maximum. Transformed into a 1-byte binary value for GNDB storage using a table of legal codes. As with Data Source Name, aliases should not be used, and the code should agree with the dictionary.

- Position: consists of two fields, latitude and longitude. Each is a signed numeric string of 7 and 8 digits (including sign), respectively. The form will be ± ddmmss ± dddmmss where positive indicates north or west and negative indicates south or east, dd or ddd are 2 or 3 digits representing degrees, mm are 2 digits representing minutes, and ss are 2 digits representing seconds. While internal GNDB representation need not be in degrees, minutes, and seconds, it is recommended. An accuracy of 1 second should be sufficient for all GNDB entries, provided that city maps are not supported by the GNDB.

- Positional Accuracy: unsigned floating point numeric, range 0 to 32,000 km. Precision to 100 meters (1-decimal digit). This is the error (believed) possible in the position, which depends on the geodetic control of the original map and other factors. It seems unlikely that precisions better than 100 meters would be needed for any unclassified geographic name.

This entity allows the cartographer to identify those geographic names not positioned with sufficient accuracy to be automatically placed on a map.

- Feature Designator: alphanumeric, 6 characters maximum. Stored in the GNDB as a 2-byte binary value, transformed for storage based on a table of legal codes. This entity indicates to what sort of object the geoname applies. Codes must be consistent with the data dictionary.

- Feature Attribute: unsigned 6-digit numeric. Could indicate population for a population center (in thousands of people), size of a waterway, etc.

- Administrative code: binary, 1 byte. Indicates if a population center has special administrative significance (e.g., capital of province, state, country).

- Acceptance Flag: binary, 1 byte. Indicates whether or not the name is approved (by the Board on Geographic Names or DMA).

- Province or State: alphanumeric, 20 characters. If the feature lies within some administrative region within the country, this field allows the AADES to specify the inclusion. The name of the province or state should also be entered in a separate data record, giving its attributes. It would be preferable to rank entries, entering the geographic names associated with these provinces or states before giving the geographic names of features within them.

- Bounding Rectangle: five fields, two latitudes, two longitudes, and a binary field (total of 13 bytes). Each latitude or longitude is a signed numeric string, either +ddmmss or +dddmmss, and the binary field is 1 byte. The GNDB will not contain the boundary data, but for area retrievals it will contain the four coordinates of a latitude-longitude "rectangle" that contains the feature. The first four fields give the top, bottom, left, and right edges of this rectangle. This data is digitized only when the feature lies wholly within the map or extends off the map on one or more sides (Fig. 5.1). The binary field contains four bits, which indicate whether the feature extends off the top, bottom, left, or right of the map.

- Non-Anglicized Name: alphanumeric, 40 characters with provision for overflow. This field would be used primarily for NATO countries. The name would be in Latin alphabet, upper or lower case, with diacritics. If the name did not fit into 40 characters, it would be handled by an overflow to the next record.

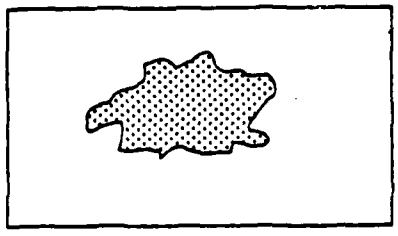
- Area Code: alphanumeric, 4 characters. Depends on the position of the feature. It is derived from position and country by GNDB applications software.

- UTM Grid: alphanumeric, 8 characters. Currently fewer characters are used, but 8 allows more accuracy and flexibility. UTM grid coordinates are computed by applications software in the GNDB.

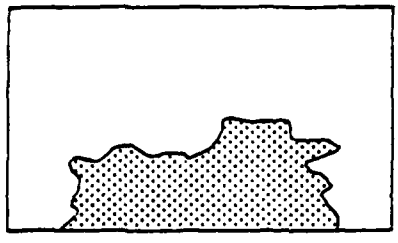
- Selected Map Sheet: alphanumeric, 7 characters. For a particular product, a map scale of 1:50,000, 1:100,000, 1:200,000, 1:250,000, 1:500,000, 1:1,000,000, 1:2,000,000, 1:3,000,000, or 1:5,000,000 may be chosen, and the map sheet of that scale that contains the position of the feature will be computed and placed here.

- Alias Name: alphanumeric, 40 characters. Each name may have to up 6 aliases. Multiple aliases may be treated with multiple records. Overflow is handled as for other names.

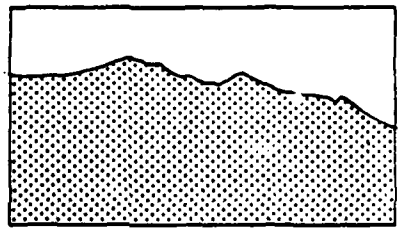
- Territorial Name: alphanumeric, 40 characters maximum. Territories, as far as the data base is concerned, are regions administered by a country that the user may not want to



a) Feature wholly contained in map



b) Feature extending off one side of map



c) Feature extending off several (three) sides of map

Figure 5-1. Relations of Areal Boundaries to Map Limits

consider when making a query on the country. For example, a user working on a map of Europe may call up Spain in the data base, but not be interested in Melilla, since it is in Africa. As in the case of country name, it is allowable to have aliases and abbreviations, but each territory must have a unique "official" territorial name.

- Index of Geonames: binary, 1 byte. To establish a primary key in the data base, an internal index is associated with a geonames and country. The index is sequentially assigned so that the triple (geoname, index, geopolitical entity) is unique within the data base. This triple is called the geoname key.

- Global Area Name: alphanumeric, 20 characters. Global area names refer to collections of geopolitical entities (e.g., "Soviet bloc") as an aid to the interactive user. A catalog of legal names will be in the data dictionary.

5.2 Data Sets

5.2.1 User Code Files

GNDB users will have identification codes and passwords. Identification codes are open codes used to log onto the system, to identify other users on the system, to route system mail, and to identify

the parties responsible for various entries and changes in the data base. Identification codes are alphanumeric, with a maximum length to be established at implementation. This code should be at least 3, preferably 6, and not more than 8 characters long.

The password is a protected code, known only to the individual user and the system manager. Access authorization will be withdrawn when the user is no longer assigned to a job requiring GNDB access, but identification codes of all operators with update authority will be maintained to facilitate audit trail processing, which includes user identification codes associated with entry modifications.

To reduce storage in the audit trail data file, the alphanumeric identification codes will be translated into binary indices (see Section 7). Although 1 byte would provide adequate space initially, 2 bytes should be allocated to these binary indices for future growth.

For each active user code, access authorization levels will be established for various files. Access authorization may be at three levels:

- no access to file,
- read only access,
- read and write access.

Each active user will be assigned the appropriate level of access to:

- GNDB,
- auxiliary files,
- private files (other than one's own),
- management information statistics files,
- operation system files.

The user code file provides a structure to handle these data entities. Password and identification code features will be implemented as part of the operating system. The file access codes may be implemented either as part of the OS or as part of the DBMS. Binary indices may be implemented as part of the data dictionary or as part of the DBMS.

5.2.2 Audit Trail Data File

This file maintains temporal information concerning the data stored in the GNDB. As a logical subunit of the GNDB, its contents should be available for query along with GNDB data elements. In terms of physical structure, this file will probably be on a storage unit different from the geonames data and will be accessed less frequently than GNDB data entities.

Contents are logically divided into two groups: origination dates and audit groups. For storage, dates are transformed into Julian day counts. A field length of 2 bytes can represent the period 1880–2058, which should be adequate for both old data sources and future system needs.

- Origination Dates: binary, 2 bytes. For each record of a relation in the GNDB, there will be an origination date, the "date of birth" of the record within the GNDB. This will apply to:

  - country alias relations,
  - territorial relations,
  - geoname data relations,
  - geoname alias relations,
  - inclusion relations.

- Audit Group: binary, 7 bytes. An audit group consists of

  - data source name: binary, 1 byte, translated from the 10-character alphanumeric data source name field;
  - date of data source: binary, 2 bytes, translated from the 7-digit numeric string;
  - date of last update (or date of data capture, if new): binary, 2 bytes, the encoded date of modification or capture.
  - user identification code: binary, 2 bytes, the binary index derived from the alphanumeric identification code.

An audit group will be applied to data in the geoname data relation. Applying the audit group at the record level gives insufficient information to track detailed changes, while applying it at the data entity level provides more information than is needed. Within each record, applying two or three audit groups to subsets of the records is desirable. An example of record partitioning into audit groups is shown in Table 5.1. The functional requirement is for the level of detail in the audit trail to be appropriate for supporting DMA objectives; the appropriate level is not yet established.

Table 5.1 Partitioning a Names Data Record into Audit Groups

Audit Group	Data Entities
1	Geoname Geoname Index Geopolitical Entity Position Positional Accuracy Feature Designator
2	Feature Attribute Military Attribute Administrative Code Bounding Rectangle
3	Type of Romanization Non-Anglicized Name Acceptance Flag Unnamed Feature Flag

5.2.3 Names Data File

A Names Data File is a toponymic workfile compiled from the GNDB at an ASP workstation. It is the sole interface between ATP and the GNDB, and the batch interface between ASP and the GNDB.

5.2.3.1 Header

The header denotes file contents and format (both are specified when the file is created). Required are

- data fields and format,
- compilation date,
- date of each use and the analyst involved,
- comments.

5.2.3.2 Contents

Names Data Files compiled from the GNDB are sets of requested geonames data for a given area and can include one or all of the GNDB entities.

5.2.4 Data Base Update File

Data Base Update Files are formatted to expedite routine toponymic comparisons and merging of same-area names data sets gathered from different sources, including those compiled from the GNDB. When discrepancies are reconciled and new information is added, the files undergo supervised entry to the GNDB.

5.2.4.1 Header

The header describes the status of toponymic research for the file as a whole. It is a 256-byte character record that includes

- analyst(s),
- file creation date,
- comments.

Source information is contained in each individual names record. Comments regarding the stage of research:

- fully researched and ready for merging with the GNDB,
- entirely unprocessed,
- partially processed, problems isolated to (problem area).

5.2.4.2 Contents

Contents are in Standard Data Transfer format (Table 5-2). Every second record is a 256-byte comment field.

5.2.5 Map Sheet Boundaries

This file holds the sheet lines of each map series. It is used to specify an area range by map sheet number rather than coordinate boundaries. It contains the following.

- Type of map: alphanumeric, 3 characters. Indicates what map series is used (e.g., JOG, ATM, etc.). The field size will accommodate a 3-character abbreviation for each DMA product.
- Map sheet number: alphanumeric, 6 characters. Index number to the map sheet within the series.
- Map sheet limits: four fields, 12 bytes. Consists of two latitudes and two longitudes specified by DMA to identify the map boundaries.

5.2.6 Coordinate Limits

Coordinate limits are used to detect positioning errors to ensure that a name's coordinates are within its geopolitical boundaries. The data structure is a 180 x 360 matrix, each matrix element

Table 5-2. Standard Data Transfer Record.

Entity Name	Size (Bytes)
Data Source Name	10 (1)
Number of Characters in Geoname	1
Number of Characters in Non-Anglicized Name	1
Number of Characters in Alias	1
Number of Characters in Province Name	1
Number of Characters in Country Name	1
Names (geoname, non-Anglicized name, alias, province name, country name)	140 (2)
Type of Romanization	1
Date of Data Source	3 (3)
Date of Data Capture	3
Date of Last Update	3
Position	6 (4)
Positional Accuracy	2
Feature Designator	6
Attribute	6
Administrative Code	1
Area Code	1
UTM grid	8
Selected Map Sheet	7
Approved or not Approved	1
Bounding Rectangle	13 (5)
Pointer to File Containing Feature Coordinates	8
Unused	32
	256

(1) The GNDB maintains a dictionary of legal data sources.

(2) If more than 140 characters are required, the next record is an overflow record. All names are stored in this field to substitute one large field with overflow allowances for potentially five large fields with possible overflows.

(3) Dates are numeric strings: ddmmyy.

(4) Position as currently planned is a point (the location of a point feature, the mouth of a river, or the centroid of an area feature) given as two signed numeric strings: +/−dddmmss and +/−ddmmss. Negative indicates latitude South or longitude East, positive indicates latitude North or longitude West.

(5) The bounding rectangle is high and low latitudes and longitudes, with an additional byte indicating if the bounding rectangle is incomplete due to the feature leaving the map.

corresponding to a 1° square on the earth's surface. If the 1° square contains only one geopolitical entity, the element indicates which entity. "Noncountry" is a geopolitical entity (ocean cells fall in this category). If the 1° square includes more than one geopolitical entity (e.g., includes a piece of an international border), the element indicates which geopolitical entities are involved. More than two geopolitical entities may be located within a 1° square.

5.2.7 Area Code Boundary File

Since area codes used in gazetteers are not generated from a mathematical formula, a file is required to convert latitude/longitude pairs into area codes. The data structure of this file is not yet established, but several alternatives are available. One approach is to represent area boundaries as linear features defined by sequences of points. A second approach is to use a variable resolution grid, with special provisions for positions near boundaries. From a design standpoint, this file should be capable of determining the area codes of geographic coordinates, but should not require detailed feature data on irregular boundaries (e.g., rivers) to the extent that it becomes a small feature file.

6.0 SUBSYSTEM INTERFACES

Under the anticipated system configuration (Fig. 6.1), the sole read-write GNDB access is through ASP. ATP has a read-only GNDB interface. This section describes Geonames Processing Subsystem interfaces. Interfacing the GNDB to external systems is considered in Section 6.4.

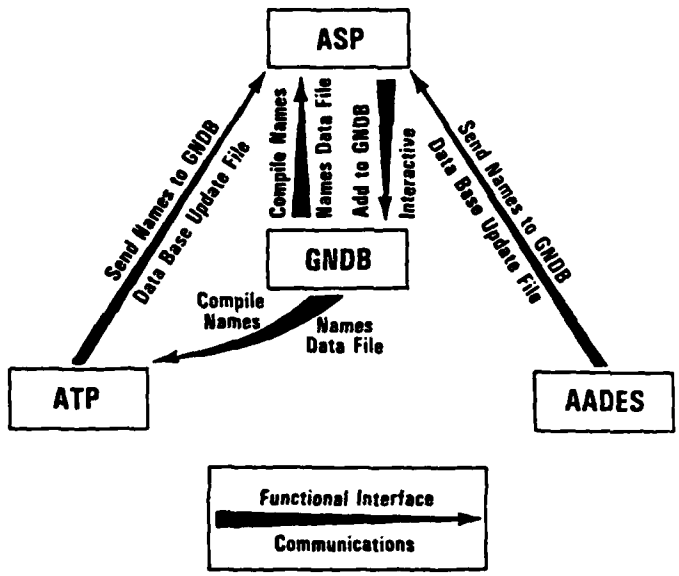


Figure 6-1. Interfaces Between the Subsystems

All interfaces use temporary data sets. One subsystem creates and names the data set, and another reads it. The user destroys the file when it is no longer needed. The interactive user interface is handled using a common GNDB-ASP work area. The interface to external computer systems is via magnetic tape or similar media.

The data entities listed are not applicable to all geographic names; likewise, information may be unavailable. There are a number of large (40-character) fields. Data structures (an implementation, not a functional matter) may have large fixed records, smaller flexible records, or even variable length records. Variable length records, however, would increase processing complexity and are not recommended.

6.1 AADES-GNDB Interface

AADES provides bulk input to the GNDB via the Data Base Update File. AADES should provide as many data entities as possible per geoname, limited only by input source detail. Table 6.1 shows the data entities that will pass from AADES to the GNDB. Entities required for every geoname are indicated by an "R" in the right column. If an optional data entity has no value, the field should contain either blanks or a zero.

6.2 ASP-GNDB Interface

6.2.1 Toponymic Updates

Initially, product generation from the GNDB is assumed to include toponymic research, verification, and updates of the contents of the GNDB since it is unlikely that full coverage will be achieved in less than 10 years. Thus, product generation will begin with compilation of all stored names data in the study area to be interactively merged and verified against Data Base Update Files. When the completed and corrected set of area geonames is defined, the required entities are written to a product file (i.e., Gazetteer File or Names Data File). Changes to the GNDB are written to a Data Base Update File.

6.2.2 Gazetteer Production

A file for gazetteer production in an area that is known to be toponymically up to date will require the data entities given in Table 6.2. Almost all entities are required, indicated by the "R" in the right column.

6.2.3 Interactive Users

Interactive GNDB use will be via ASP. Thus, this interface must handle queries, retrieve data sets, and deal with any data item in the GNDB or its associated files (e.g., map sheet files). One cannot specify a format for such an interface. Rather, this interface will be fixed length blocks containing text by which the user communicates with the DBMS software.

6.3 ATP-GNDB Interface

Initially, map production will be preceded by toponymic updates due to incomplete GNDB coverage. In this case the completed and corrected set of area geonames is run through interactive map names selection algorithms, and the required entities are written to a Map Data File. Changes to the GNDB are written to a Data Base Update File.

A file for map production in an area known to be toponymically up to date should contain the data entities listed in Table 6.3. Required entities are indicated by an "R."

6.4 Interfaces to External Systems

Although there is no stated need for the GNDB to interface directly with any system outside the Geonames Processing System, it is conceivable that such an interface may be required at some time. Any such interfacing will be done using the common format of the DMA Feature File (DMAFF) described in Ref. C. An application program will reformat GNDB retrieval files.

Table 6.1 Data Entities for AADES-GNDB Interface

Country	R
Data Source Name	R
Date of Data Source	R
Date of Capture by AADES	-
Geographic Name	R
Type of Romanization	-
Position	R
Positional Accuracy	R
Feature Designator	R
Feature Attribute	-
Administrative Code	-
Acceptance Flag	-
Province or State	-
Bounding Rectangle	-
Non-Anglicized Name	-

Table 6.2 Data Entities for Gazetteers

Geographic Name	R
Position	R
Feature Designator	R
Area Code	R
UTM Grid	R
Selected Map Sheet	R
Alias Name	R
Acceptance Flag	R
Date of last update to record	-

Table 6.3 GNDB Data Entities for Names Placement

Country	-
Geographic Name	R
Position	R
Positional Accuracy	R
Feature Designator	R
Feature Attribute	R
Administrative Code	-
Pointer to Boundary Data	R
Non-Anglicized Name	-
Acceptance Flag	-
Date of Capture by AADES	-
Date of Last Update	-

7.0 DESIGN CONSIDERATIONS

The Geographic Names Data Base must manage a massive volume of data yet provide efficient storage and retrieval facilities. Hardware and software alternatives must be examined in light of trade-offs between functional parameters to balance system capabilities. Key design parameters are data storage efficiency, system response times, ease of applications development, the level of user support expected, and the flexibility of data structures to adapt to changes. The relative importance of each parameter has a major impact on design, since improvement in one functional area may be at the expense of other system capabilities.

7.1 Architectures

The GNDB can be a single, very large data base, a data base distributed across more than one computer, or a collection of independent data bases. DBMS and hardware requirements will partially depend on the functional requirements imposed by the chosen architecture.

7.1.1 Distributed vs. Centralized Hardware

Because of the large data base size, a centralized, single-computer approach would place heavy performance requirements on both computer processing and disk drive I/O, while data base administration and maintenance functions represent a technical risk. Dividing the data base into several discrete hardware components offers improved performance and fault-tolerance, but greatly increases the complexity of maintaining consistency and administrative control of the system.

Hybrids of these basic hardware alternatives are also feasible. The capacity to download processing functions from the computer to intelligent terminals or workstations would provide additional configuration options. Maintaining centralized control over data integrity requires global administrative hardware functions.

All GNDB disk and computer CPU hardware is expected to be maintained at a central site to facilitate administration. Peripherals (terminals and output devices) may be distributed locally.

7.1.2 Multiple vs. Single DBMS

The software to control data manipulations and manage files could be centralized in a single system or networked between multiple processors or data bases. Separate DBMS modules could be operated and maintained independently, but would require global software to provide management information statistics and to assist operators in locating specific information. Commercial DBMSs are only beginning to cope with the difficulties of distributed data base management; the software technologies for concurrency control, automatic navigation, and access algorithms in a distributed environment are not mature.

GNDB data sets, however, can be partitioned to improve retrieval efficiency. Dividing data elements by country is a natural extension of current processing methodologies, since many geographic name products and queries orient on political boundaries. The increased processing efficiency of partitioning the data base by country is adequate compensation for the consequent increase in operational and administrative complexity.

7.1.3 Data Base Structures

Relational DBMS structures are well suited to query-oriented, multi-key retrieval of small data sets, but storage and processing efficiency suffers with large volumes of data. A hierarchical or net-

work structure, on the other hand, provides more efficient storage and retrieval at the cost of flexibility in data relationships and attribute sets. While the elasticity of the relational approach is desirable, performance requirements (Section 9) suggest that a hierarchical physical structure will be required.

Although the large data volumes and well-defined relationships anticipated in the Geographic Names Data Base suggest a hierarchical structure, toponymic queries on data subsets could benefit from a relational approach for identifying lexical or linguistic relationships among data elements. Several commercial DBMSs currently incorporate aspects of both relational and hierarchical structures, using hierarchical storage implementations that support relational constructs for queries.

7.2 Data Structures

The GNDB is primarily concerned with storage and transfer and requires a variety of data structures. These may vary from simple files to the GNDB itself. This section discusses GNDB logical data structures. The physical design (i.e., how data is actually stored and retrieved) is implementation dependent and, thus, is not specifically addressed. File structures are discussed in Section 5.2.

7.2.1 Interface Files

GNDB file structures that transfer data to or from other Geonames Processing subsystems, and an input file format, were discussed in Section 5. These data structures are simple files and should be implemented with fixed-length records.

7.2.2 Existing Names Files

DMA currently has some geonames data in computer-readable form, such as the Multiset III files (on diskettes, transferred to 9-track tape), the NIS files (still a prototype system), and the old gazetteer tapes (7-track, BCD). The data formats for these files are already established. For the GNDB to use them as inputs, applications programs are required to reformat the data into input file format.

7.2.3 GNDB Relations

The logical structure of the GNDB may be described using relational constructs.

- Country Alias Relation: each country name is associated with a unique "official" country name.
- Territorial Relation: each territorial name is associated with a unique "official" territorial name. Each "official" territorial name is associated with a unique "official" country name.
- Adjacency Relation: geopolitical entity names form a relationship when they are geographically adjacent.
- Global Area Relation: each global area name is associated with a non-empty list of geopolitical entity names.
- Geopolitical Entity Inclusion: each geographic name must be associated with a geopolitical entity name and a geoname index to form the geoname key.
- Geoname Alias: pairs of distinct geoname keys may be formed when two geonames are aliases for the same feature. In general, alias geoname keys in a relation will have the same geopolitical entity.

- Inclusion Relation: pairs of distinct geoname keys may be formed when one geoname is a part of the other (e.g., a town in a province, a mountain in a mountain range). In general, both geonames in a relation will have the same geopolitical entity.

- Geoname Data Relation: The triple (geoname, index, geopolitical entity) is the primary key to this relation, which links geoname, geoname index, geopolitical entity, position, positional accuracy, feature designator, feature attribute, military attribute, administrative code, bounding rectangle, type of Romanization, non-Anglicized name, acceptance flag, and unnamed feature flag.

8.0 ASSUMPTIONS AND CONSTRAINTS

This functional design specification uses the best estimates of conditions and policy. However, changes may occur. This section describes the assumptions incorporated in the report.

8.1 Security

This report assumes that the GNDB will be unclassified. If a supplemental classified base is built, it should use its own operating system, and all terminals in unclassified areas should be denied access to this supplementary system. The distributed data base architecture discussed in Section 7.1 lends itself to classified operations. However, classified data must be stored on demountable, not fixed, disks.

8.2 Hardware

This report assumes that the Geonames Processing System will not be required to use DMA's current Univac equipment. Instead, hardware alternatives should be considered according to functional requirements and cost. A restriction to Univac hardware would impose serious limitations on data base selection and performance.

8.3 Feature Data

An assumption is made that the GNDB will not hold detailed feature boundary data but that such data can be obtained from appropriate feature data bases.

8.4 Non-Romanized Names

An assumption is made that the data base will contain non-anglicized but not non-Romanized names.

Non-Romanized names can be included using one of two devices. First, they can be stored as bit maps (images), requiring only the addition of a character data entity (length to be determined). These bit maps could be accessed and displayed through their Romanized counterparts. The bit map strategy is favored if a strong non-Romanized requirement exists. The reasoning for this recommendation is discussed in Volume 5, Appendix C, of this series.

Alternatively, ASCII-coded non-Romanized names could be stored by adding a 1-byte data entity (the Alphabet Translation Table) to indicate in which alphabet the encoded name should be displayed.

8.5 Time Frame

An assumption is made that the GNDB will be built over a period of time as part of the normal production process. If a high-intensity effort is chosen, to load the data base as quickly as possible, hardware requirements would increase to accommodate heavy AADES-GNDB traffic and massive quality control processing.

8.6 Location

An assumption is made that the GNDB will reside at DMAHTC. All updating and modification to the base will be done by DMAHTC users. DMAAC may request data from the GNDB, but would not be involved in building or maintaining the base. If a copy of the GNDB is placed at DMAAC, procedures must be formulated to ensure consistency of data base versions, concurrency of modifications, and overall data base control.

9.0 DBMS PERFORMANCE REQUIREMENTS

As the interface between the stored data files and the applications programs, the data base management software must protect the data and expedite access and retrieval. Critical aspects of this function, along with measures of retrieval efficiency, are described in this section.

9.1 Data Integrity

A data base is useless if the accuracy or validity of the stored data is questionable. The DBMS should allow definition of data elements both by type (e.g., integer, character string) and by range limits of acceptable values, and have facilities to allow applications programmers to impose additional validation checks as required. Hardware or data transmission errors should be detected and flagged, and failures during a processing sequence should be rolled back to the last correctly processed record with appropriate error messages. Data relationships (e.g., parent/child, one-to-many) established by the DBA should be protected from unauthorized modification, and mandatory relationships should be enforced.

9.2 Physical Data Protection

The volume of data in the GNDB will make frequent back-up copying infeasible. The DBMS should provide update facilities that do not require manual re-entry of changes for both "father" and "grandfather" back-up copies in the event of media or storage device catastrophe. Accurate records of physical media storage contents should be maintained. System facilities for protecting data are also addressed in Section 2.3.

9.3 Data Security

The DBMS must prevent unauthorized users from modifying data or reading sensitive data. Access restrictions to the level of data item (not just to files or records) are required. System security administration was addressed in Section 2.2.

9.4 Data Independence

To preserve flexibility for future enhancements, the DBMS should provide data independence between applications software and the physical structure of the data base. It should be possible to modify the physical structure without affecting either the logical structure (the user's view of data base) or previously written data access programs.

9.5 Storage Efficiency

Controlled redundancy of data elements is required to support data base consistency and integrity and to reduce storage requirements. Variable length record processing and automatic data compression is desirable. The DBMS must support data distributed across multiple storage units, and multiple users should be able to share common memory buffer space.

9.6 Retrieval Efficiency

The DBMS should provide a variety of access methods (e.g., hashing, indexed pointer arrays) to optimize retrieval. Facilities to provide data migration or data clustering (physical proximity of records that are frequently accessed in conjunction) are desirable, along with flexibility in designating (or accepting from the operating system) various page and buffer memory sizes. In short, the DBMS should support control of the GNDB physical structure.

Requests for information from the data base can be categorized in three time frames:

- interactive—small exploratory queries, quality control input resolution, and some system status statistics;
- short-term—table compilation, complex queries, and management information statistics; and
- batch—input processing, Names Data File compilation, and audit trail documentation.

System response times for interactive requests should not exceed 30 seconds; for short-term requests not more than 4 hours (80% within 2 hours); and for batch runs not more than 12 hours, with 80% processed in 6 hours or less. Batch processing will generally be reserved for overnight or other slack times to improve interactive system performance.

Provisions for periodically measuring retrieval efficiency are desirable in the DBMS, and alternative data storage organizations should be benchmarked or simulated prior to implementation. The following five ratios (adapted from Reference H) measure DBMS retrieval performance in the range (0, 1), with 1 indicating a high level of efficiency.

$$(1)\ \text{Recall ratio} = \frac{\text{number of relevant record retrieved}}{\text{number of relevant records stored in data base}}$$

This measure would be useful in testbed trials to verify the adequacy of proposed access methods when the desired number of successes is known beforehand.

$$(2)\ \text{Precision ratio} = \frac{\text{number of relevant records retrieved}}{\text{total number of records retrieved}}$$

This measure tests DBMS discrimination of defined entity sets.

$$(3)\ \text{Logical efficiency} = \frac{\text{number of relevant data elements}}{\text{total number of elements references}}$$

This ratio measures the efficiency of the data retrieval algorithm in terms of access path lengths. The total number of elements referenced includes index and pointer elements.

$$(4)\ \text{Physical efficiency} = 1 - \frac{\text{total number of blocks or pages accessed}}{\text{number of relevant records retrieved}}$$

This measures the efficiency of I/O operations, often a major source of bottlenecks in retrieval.

$$(5)\ \text{Data efficiency} = \frac{\text{number of relevant characters or bits retrieved}}{\text{total number of characters or bits accessed}}$$

This evaluates how much unnecessary data is processed to answer sample queries. Clustering techniques can improve this ratio for common query types.

10.0 HARDWARE REQUIREMENTS

The GNDB requires enough computational power to support 20 interactive terminals operating in a DBMS environment. The throughput performance of each hardware component required to produce acceptable system response times will depend on the configuration selected, since disk storage size and timing, CPU memory, and local memory in intelligent terminals can be traded off to produce similar system capabilities.

The degree of integration between system components, the operating system, and the DBMS is relevant to hardware selection. Since OS capabilities generally depend on what features are supported by a particular hardware vendor, care must be taken to examine the compatibility of available operating systems with alternative DBMSs. Added to system performance considerations are hardware cost/efficiency; compatibility with other subsystem equipment; reliability; equipment cost and leasing options; vendor assistance, support, and reputation; delivery and installation requirements; and maintenance facilities.

10.1 Mass Storage

The eventual GNDB size makes efficient input/output between the CPU and on-line storage critical. The estimated 4.25 x 109 bytes of data (see Reference A, p. 131) will require several disk drives to physically store the data on-line. Since the data base will not be classified, removable disk media are not functionally required. Recent advances in fixed-disk technology have provided storage capacities, reliability, and average access times superior to removable disk drives. Storage devices should provide at least 500-MB capacity, an average access time (see time plus rotational delay time) of not more than 40 msec, and a data transfer rate of at least 5 Mbits/sec.

10.2 CPU

Although only a small portion of the data base will be resident in memory at one time, storage and control of 20 concurrent operations require significant memory and processing. A single-CPU architecture should have a minimum of 4-MB core memory, or 2 MB with virtual memory capabilities. A design linking more than one CPU to share data base processing reduces per-CPU memory requirements, but entails additional system overhead to coordinate global processing. The sum of multiple CPU memories should be at least 1.2 times the memory requirements of a single-CPU design.

CPU cycle time or processing speed (instructions per second) are not critical to system throughput, which will be I/O bound rather than CPU-intensive. In general, processing capabilities are proportional to CPU cost. Relevant evaluative criteria include the number, type, and capacity of I/O data channels supported; provisions for memory and peripheral expansion; and special features such as parallel operations or multiple processors, which may significantly enhance DBMS capabilities.

10.3 Peripherals

The GNDB requires a variety of input and output devices, including tape drives, work stations (CRTs and keyboards), printers, and digital plotters. Additional input devices such as mice, touch-screen displays, and digitizing tablets may facilitate quality control processing and edit/format routines. Final specification of the number and type of each kind of peripheral must be made in light of Geonames Processing System requirements rather than in terms of the GNDB, since hardware requirements overlap subsystem boundaries.

The principle functional requirement for peripherals is compatibility with system software and hardware protocols, instruction sets, and transmission rates. Ergonomic and performance requirements are described in Volume 5 of this series.

Approximately 20 interactive ASP workstations are needed for toponymic and cartographic data base queries, and several terminals for input quality control and programming support are desirable. Interactive CRT capabilities are addressed in Section 3.2.

At least two tape drives are required to support the AADES/GNDB interface. Additional drives for data base backups and audit trails are desirable. Two high-speed printers are needed to spool batch output; one medium-speed printer (150–300 cps) for each of four workstations is desirable to generate draft gazetteers and map name tables; and low-cost printers with screen-dump facilities for each workstation are desirable to minimize operator paperwork. Several multicolor pen plotters are desirable to facilitate quality control and cartographic queries.

APPENDIX

REFERENCES

A. Brown, R., et al., *Advanced Type Placement and Geonames Database: Comprehensive Plan*, NORDA Technical Note 189, January 1983.

B. Chorafas, D. N., *Databases for Networks and Minicomputers*, Petrocelli Books, 1982.

C. Defense Mapping Agency, *Specifications for Prototype DMA Feature File (Draft)*, November 1983.

D. Hubbard, G. U., *Computer-Assisted Data Base Design*, Van Nostrand Reinhold, 1981.

E. Inmon, W. H., *Effective Data Base Design*, Prentice-Hall, 1981.

F. Monmonier, M. S., *Computer-Assisted Cartography*, Prentice Hall, 1982.

G. Shah, A.D., "Data Administration: It's Crucial," *Datamation*, January 1984, pp. 187-192.

H. Wong, P. M. K., *Performance Evaluation of Data Base Systems*, University Microfilms International, 1981.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A162504

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS None
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NORDA Report 99	5 MONITORING ORGANIZATION REPORT NUMBER(S) NORDA Report 99
6 NAME OF PERFORMING ORGANIZATION Naval Ocean Research and Development Activity	7a NAME OF MONITORING ORGANIZATION Naval Ocean Research and Development Activity
6c ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004	7b ADDRESS (City, State, and ZIP Code) Ocean Science Directorate NSTL, Mississippi 39529-5004
8a NAME OF FUNDING SPONSORING ORGANIZATION Defense Mapping Agency 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code) HQ/STT Washington DC 20305	10 SOURCE OF FUNDING NOS PROGRAM ELEMENT NO: 64710B; PROJECT NO; TASK NO; WORK UNIT NO

11 TITLE (Include Security Classification) The Geonames Processing System Functional Design Specification, Volume 2: Geographic Names Data Base

12 PERSONAL AUTHOR(S)
Gail Langran, Allen Barnes*, and Steven Miller*

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Yr., Mo., Day)	15 PAGE COUNT
Final	From _____ To _____	March 1985	37

16 SUPPLEMENTARY NOTATION
*with Planning Systems, Inc., McLean, Virginia

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB GR	Maps, computers, software systems

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This report describes the Geonames Processing System attributes and serves as a basis for understanding between the user and the developer. The subsystems referred to are: Advanced Symbol Processing, Advanced Type Placement, Geographic Names Data Base, and Automated Alphanumeric Data Entry System.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ■ DTIC USERS ☐	Unclassified	
22a NAME OF RESPONSIBLE INDIVIDUAL Gail Langran	22b TELEPHONE NUMBER (Include Area Code) (601) 688-4449	22c OFFICE SYMBOL Code 351

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

END

FILMED

1-86

DTIC